The Treatment of Missing Data on Placement Tools for Predicting Success in College Algebra at the University of Alaska

by

Alyssa Crawford

A

PROJECT

Presented to the Faculty

of the University of Alaska Fairbanks

in Partial Fulfillment of the Requirements

for the Degree of

MASTER OF SCIENCE

By

Alyssa Crawford, B.S.

Fairbanks, Alaska

May 2014

0pt

Abstract

by Alyssa Crawford

This project investigated the statistical significance of baccalaureate student placement tools such as tests scores and completion of a developmental course on predicting success in a college level algebra course at the University of Alaska (UA). Students included in the study had attempted Math 107 at UA for the first time between fiscal years 2007 and 2012. The student placement information had a high percentage of missing data. A simulation study was conducted to choose the best missing data method between complete case deletion, and multiple imputation for the student data. After the missing data methods were applied, a logistic regression with fitted with explanatory variables consisting of tests scores, developmental course grade, age (category) of scores and grade, and interactions. The relevant tests were SAT math, ACT math, AccuPlacer college level math, and the relevant developmental course was Devm/Math 105. The response variable was success in passing Math 107 with grade of C or above on the first attempt. The simulation study showed that under a high percentage of missing data and correlation, multiple imputation implemented by the R package Multivariate Imputation by Chained Equations (MICE) produced the least biased estimators and better confidence interval coverage compared to complete cases deletion when data are missing at random (MAR) and missing not at random (MNAR). Results from multiple imputation method on the student data showed that Devm/Math 105 grade was a significant predictor of passing Math 107. The age of Devm/Math 105, age of tests, and test scores were not significant predictors of student success in Math 107. Future studies may consider modeling with ALEKS scores, and high school math course information.

Contents

1. Introduction

Research in the initial placement requirements for math courses taught at the University of Alaska (UA) are important to ensure student success. A missing data method is necessary for the analysis of student placement for Math 107, an entry-level math course on college algebra at UA. The placement system was flexible because a common placement test was lacking until recently. Students may have used up to four different placement mechanisms including developmental coursework and tests scores. Many students placed into Math 107 by having a satisfactory grade in a developmental course titled intermediate algebra (Devm/ Math 105). Also, students may have used AccuPlacer college level math score for placement into Math 107 and historically, students have used SAT Math or ACT Math scores. More prerequisite information of the UA system and its three universities may be found in the appendix tables A.1, A.2 and A.3. This project addresses the issue of missing data using two missing data methods to answer the following question:

> Are SAT, ACT, AccuPlacer and MATH/DEVM 105 score/grade and age significant predictors of successful completion of Math 107 for baccalaureate degree seekers?

Age refers to the age of test or coursework and is defined as the time elapsed between the start of the first attempt in Math 107 and the test date, or end date of coursework.

This project compared and contrasted two missing data techniques: complete case deletion, and multiple imputation. Complete case deletion is a common method in which any observations with missing information are deleted. Multiple imputation produces several data sets, each with different reasonable values filled in for the missing data, and the data sets are combined for an overall analysis. For this

project, the method of multiple imputation was implemented by the R package Multivariate Imputation by Chained Equations (MICE). It is important to specify the R package since several R packages exist for multiple imputation and the R packages use different methods of imputation.

After applying the missing data method, a logistic regression model was fitted. The model follows,

$$\begin{aligned} y_i \sim & \text{Bernoulli}(\pi_i) \\ \text{logit}(\pi_i) = & \beta_0 + \sum_{j=i}^{J} \beta_j x_{i,j} \end{aligned}$$

where $\pi_i = P(y_i = 1)$ is the probability of success. Success is defined as a final grade C or above for the first attempt of Math 107. We let $X_{i,j}$ denote the jth explanatory variable ($j = 1, \ldots, J$) for the ith student. Explanatory variables were scores and ages of ACT, SAT, AccuPlacer, and grade and age of Devm/Math 105. Interaction between age and score or grade were included when possible. The regression coefficients (β's) explained the relative effect of the explanatory variable. This project was interested in unbiased estimation of the β's, especially after applying the missing data technique.

This project used readily available information from UA's Decision Support Database that included system-wide information from UA administrative information systems. Student registration information between Summer, 2007, and Spring, 2012, revealed that 4,793 students met inclusion criteria for the project. The general criterion was first attempt at taking Math 107, among 4-year degree seeking students. Other details may be found in the methods section inclusion and exclusion criteria or Figure B.1.

In order to choose between the complete case deletion or multiple imputation methods for the student data, we conducted a simulation study to investigate the two methods under a high degree of missingness and correlation. The best method is the one with the least biased estimators and good confidence interval coverage.

1.1 Missing Data Background

1.1.1 Missing Data Patterns

Any discussion of missing data begins with Rubin's (1976) classification of missing data patterns: missing completely at random (MCAR), missing at random (MAR), missing not at random (MNAR). These patterns are simply mathematical devices to describe why values are missing, and are included here to help describe how well a missing data technique will perform. The goal of missing data procedures should not be to make estimates of missing data but to make valid inferences (Schafer and Graham, 2002). Usually the goal is not to re-create the lost data but rather to produce unbiased estimators (van Buuren, 2012). The mechanisms - MCAR, MAR, and MNAR - are helpful in describing how estimators will perform and the following definitions are conceptional.

- Missing Completely at Random (MCAR) occurs when the probability of missing data on a variable is independent of other measured variables, and of the value itself. An example could be a student who was sick on the day the SAT test was given.
- Missing at Random (MAR) occurs when the probability of missing data on one variable is related to other measured variables but not on the value itself. An example of MAR would be a student who didn't take an AccuPlacer test because of their high SAT or ACT Score.
- Missing not at Random (MNAR) occurs when the probability of missing data depends on the value itself. An example could be a student who didn't submit a low SAT score to the University because it was low.

The above three missing data patterns may inform us of how a missing data technique will perform and this is discussed in the next section.

1.1.2 Missing Data Techniques

Missing data techniques are categorized as either traditional or modern. Traditional techniques include complete case deletion and single imputation, while modern techniques include multiple imputation and the EM algorithm.

Complete case deletion, also called listwise or case wise deletion, is a common technique implemented by most statistical software. For this technique any observation with a missing value is thrown out. For MCAR, there is general agreement that deletion of missing cases leads to no bias in estimators, but at a cost of reduced sample size (Baraldi and Enders 2010). A smaller sample size leads to greater standard errors and reduced power. However, when the MCAR assumption is not valid, deletion of observations produces biased estimators (Baraldi and Enders 2010). The traditional technique of complete case deletion performs well when the probability of missing values on one variable are independent of other variables or the value itself (MCAR) but the technique performs poorly when probability of missing data on one variable are correlated with other variables (MAR).

Single imputation means filling in missing data with reasonable values and results in one complete imputed data set. Reasonable values could be the average of all the observed data, or a fitted value from a linear regression on other variables. Single imputation was not used here due to general agreement that single imputation leads to biased estimators under the assumptions of MCAR or MAR (Baraldi and Enders, 2010). It also seriously underestimates the vairance of the cofficents (β's).

Modern data missing techniques like multiple imputation and maximum likelihood (EM Algorithm) are far more complicated techniques that are not perfect fixes to missing data, yet are highly recommended since they produce unbiased estimates under MAR (Baraldi and Enders, 2010). Multiple imputation is the main focus in this project since software for maximum likelihood like the EM algorithm is not yet implemented for logistic regression models with missing data. For multiple imputation, several data sets are created each with different imputed values and these data sets are combined for analysis. Multiple imputation is best explained in three steps: imputation, analysis and pooling. Multiple imputation involves first an imputation step where several copies of datasets, each with different imputed values, are produced. The method of imputation depends on software, and the most important aspect of the imputation step is defining a imputation model. For the second step, each dataset is analyzed separately using the same statistical model. The final step is to pool together estimators using rules from Rubin (1987). There are two estimates of variances: one for sampling variance and the other describing extra variance caused by missing data. Of these steps, the imputation stage is the most difficult, due to tough decisions about the imputation model.

1.1.3 Placement Background

Several barriers exist when analyzing student placement information because of the historically inconsistent and flexible placement polices across the Universities. Each University campus may have had different cutoff scores for the same Math course and a common placement test was lacking until recently. Another barrier is the lack of readily available official record of a specific prerequisite for a student. This information may exist in an official capacity but is not available for efficient analysis or development of management information. Also, sometimes students have instructor permission to enter a course, and there is no official record for reason. It appears that enforcement of each University's policies depends on the instructors, and instructors vary on the degree of enforcement. The universities seem to show a reliance on tests in determining placement of students.

Additionally, we briefly make a case that student high school information is likely an important predictor of student success. A study revealed that placement tests AccuPlacer and Compass were weakly associated with college GPA while high school GPA is strongly associated with college GPA (Belfield and Crosta, 2012). Another study on universities with optional standardized testing policies for admissions showed little differences in graduation rates and cumulative college GPA between submitters and non-submitters (Hiss and Franks, 2014). The same study showed that students with strong high school GPAs tended to have strong cumulative college GPAs even with low or modest testing. These studies looked at student success at a broad level of cumulative college GPA, however the project presented here looks to investigate student success in a course. Still, it is likely that student high school math grades and intensity are important to predicting student success in Math 107. High school information was not used in this project because it was not available for efficient analysis.

2. Methods

2.1 Student Data: Inclusion and Exclusion Criteria

The following section describes who was eligible to be in the study. Please also refer to figure B.1 in appendix B for a summary.

An important aspect of this project was to create an appropriate dataset from the UA data system. This project examined UA students who were 4-year degree seeking at the time of Math 107 and who first attempted the course between fiscal years 2007 and 2012. Students included in the dataset were enrolled for credit and had an official grade in Math 107 between Summer 2007 and Spring 2012. Official grades were needed to determine the response variable pass or fail; grades included as passing are grades C and above. We note that grade C- was considered not to be passing. Final grades W and AU were assumed to be progressing toward a final course grade below C, and were considered failures. There were 92 students with final grades deferred (DF), no basis (NB), and not submitted (NS) and these were not included in the dataset because these students likely had extreme situations preventing them finishing the course with a GPA eligible grade.

We checked and obtained test age and test score for placement tests AccuPlacer College Level Math, and college-admission exams ACT Math, and SAT Math. The test age was the time elapsed between the test date and the Math 107 start date. If students had re-takes of the same test, the test with the most recent test date was chosen for the dataset. Any test taken 31 days after the start date for Math 107 was disregarded. Only 82 students who first-attempted Math 107 had on record a COMPASS College Algebra exam or ASSET exam; these exams were not used in the project.

For developmental coursework, Devm/Math 105, the age of the course was the time elapsed from last date of Devm/Math 105 to the start date of Math 107. If students had re-takes of Devm/Math 105, we selected the most recent course for the dataset. Any Devm/Math 105 taken after the first attempt of Math 107 was disregarded from dataset.

Students who were previously enrolled for Devm 066 were not included in the study; this included 12 students. Devm 055 is a UAF course titled Advanced Math Fast Track: Elementary/Intermediate Algebra Review. The course is a 20 hour review of algebra and was shown to increase pass rates greatly. Students who took Devm 055 are likely better prepared and quite different from other students.

Also excluded from the study were 388 students that didn't have any relevant tests or developmental courses that were the subject of this project and because they were likely quite different from the main student population.

2.2 Statistical Methods

2.2.1 Complete Case Deletion

For the complete case deletion, a separate logistic regression is fitted for datasets having one of the predictors Devm/Math 105, SAT, ACT, or AccuPlacer. So there were four models, each with a coefficient for score/grade, age and interaction. Let Y_i be a binary variable denoting pass or fail grade for student i $(i = 1, ..., n)$. Passing grade includes grades C and above for the first attempt in Math 107. The

model is then,

$$
\begin{aligned}
y_i \sim & \text{Bernoulli}(\pi_i) \\
\text{logit}(\pi_i) = & \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{1,i} x_{2,i}
\end{aligned} \tag{2.1}
$$

where

$$
x_{1,i} = \begin{cases} \text{continuous value test score} \\ \text{or} \\ \text{ordinal value Math/Devm 105 Grade} \end{cases}
$$

$$
x_{2,i} = \begin{cases} \text{Recent (1)} & \text{Age of SAT, ACT, or Math/Devm 105} < 2 \text{ year} \\ & \text{or} \\ & \text{Age of AccuPlacer} < 1 \text{ year} \\ \text{Not Recent (0)} & \text{otherwise} \end{cases}
$$

For the complete case deletion method, the model included one placement criterion at a time, due to the very few students having all placement criteria (83 out of 4,793).

For grades in Devm/Math 105, the letter grades A, B, C, D, and F were given values 5, 4, 3, 2, and 1, respectively. For grades with plus or minus, the value was the grade point rounded with usual rounding rules. For example a letter grade A- has grade point 3.7, and this was rounded to 4 for the model. Grades AU and W were given the same grade value as F. For grade incomplete (I), the grade value was 3.0. Explanatory variable, age, was coded as a binary variable of either recent (1) or not recent (0). The age of test or coursework was the time elapsed between the start of the first attempted of Math 107 and the date of the most recent test or end of most recent Devm/Math 105 coursework. The recent category for age included ages for SAT, ACT, and Devm/Math 105 that were less than two years and not recent if older. Since many students take the AccuPlacer test right before Math 107, it was more reasonable to have recent age be less than one year and not recent if older. The baseline for the logistic regression model was always not recent (0).

Model selection began with the most complex model and we dropped terms with the largest P-values until remaining coefficients were all significant (P-value $<$ 0.05). Also, for model selection Akaike information criterion (AIC) values were used.

For statistical inference, a description of the odds ratios is given. The magnitude of coefficients in a logistic regression model is usually interpreted as odds ratios given by $e^{\hat{\beta}}$; that is, the odds of $X = x + 1$ divided by the odds at $X = x$. For odds ratios with a categorical variable, the baseline of the logit model was not recent (0).

2.2.2 Multiple Imputation

The first concern with multiple imputation is addressing the MAR assumption; there is more discussion about the MAR assumption in the conclusion section. The imputation step of the MICE package imputes missing values using the MICE algorithm. The general process imputes on a variable-by-variable basis by specifying an imputation model for every variable while other variables are treated as explanatory variables (including the actual response variable) with no missing values. The MICE algorithm is an iterative process where the final dataset is the last imputation of several cycles of imputation; the default is 5 iterations. The steps of the algorithm are described below, but first notation must be stated. We use X to denote a n x $(p+1)$ matrix of the data, with one row for each observation, one column for each of the p-many explanatory variables and one column for the response variable. Let $i = 1, \ldots, n$ indicate the rows and $j = 1, \ldots, (p + 1)$ the columns of the X matrix. Also, we have a n x $(p + 1)$ matrix, R, with entries $r_{i,j}$=1 if $x_{i,j}$ is observed and $r_{i,j}$=0 if $x_{i,j}$ is missing. The (p+1)th column is for the actual response variable, so the entries of the p+1 column of R are 1. The X matrix has missing values, and can be partitioned, $X = (X_{obs}, X_{miss})$ where X_{miss} corresponds to 0-entries in the R-matrix. Let ϕ_j be a vector of parameters $(\boldsymbol{\beta_j}, \sigma^2)$ or $(\boldsymbol{\beta_j})$ for each column j of X, if the imputation model is a Bayesian multivariate normal linear model or a Bayesian logistic regression model, respectively. A Bayesian multivariate normal linear model is the imputation model if column j corresponds to a continuous variable such as test scores or Devm/Math 105 grade, and a Bayesian logistic regression model if column j corresponds to a binary variable such as recent versus not recent test score. For the imputation model note that we consider column j as the temporary response variable denoted as X_j and the other p-many columns (including column p+1, the actual response variable) as the explanatory variables denoted by X_{-j}. Let x_j^{miss} be entries needing imputation from column j and x_{-j}^{miss} be the subset of rows of X_{-j} for which x_j^{miss} is missing.

The algorithm used by MICE was as follows:

1. Starting imputations were random draws from the observed data. The software filled in values for the entries of X_{miss} (i.e. entries in X that correspond to locations of 0's in the R-matrix) with random draws from the observed data.

2. For each iteration (t in 1, 2, 3 ..., T) :

   For each column j of explanatory variables (j=1,..,p) with missing values:

   i. Draw vector $\boldsymbol{\phi_j}$ from its fully conditional distribution.

      A. If column j is a continuous variable then $\boldsymbol{\phi_j} = (\boldsymbol{\beta_j}, \sigma^2)$, and we used the usual formulas for the least square estimator β's and σ^2 with a small amount of added random noise (see algorithm 3.1, van Buuren, 2012).

      B. If column j is binary variable then $\boldsymbol{\phi_j} = (\boldsymbol{\beta_j})$, and we used a iteratively reweighted least square estimator with added random noise (see algorithm 3.4, van Buuren, 2012).

   ii. Simulate new values x_j^{miss} using the $\hat{\boldsymbol{\phi}}_{\boldsymbol{j}}$ as follows:

      A. If continuous predictor, then we used the formula for conditional distribution of normal random values to simulate x_j^{miss} then we added random noise, $N(0, \hat{\sigma^2}_j)$.

      B. If binary predictor, then for each missing value in that column calculate $\hat{\pi}$, the inverse logit of linear combination of predictors, and simulate a 0/1 from Bernoulli distribution.

3. At the end of the Tth iteration in (2) we recorded the imputed values that were simulated.

4. Steps 2 and 3 were done in parallel for a total of m=10 imputed datasets.

Note: A separate set of $\beta's$ are estimated for each column containing missing values. And these β's are unrelated to the β's we estimate for the eventual logistic regression that uses pass/fail for Math 107 as the response vairable.

We point out that the Bayesian normal linear model imputed all missing values for a particular column in one shot, while the Bayesian logistic regression model

imputed values one at a time within each column. For the imputation models, multivariate normal data was an assumption; this assumption is known to be robust against departures (van Buuren, 2012). The imputation models were Bayesian multivariate normal linear model for test scores, and Devm/Math 105 grade and a Bayesian logistic regression model for the categorical variables ages of tests, and ages of Devm/Math 105. There were 30 iterations of the MICE algorithm and a total of 10 imputed datasets.

For the analysis step, ten imputed datasets were fitted with the following model,

$$\begin{aligned} y_i \sim & \text{Bernoulli}(\pi_i) \\ \text{logit}(\pi_i) = & \beta_0 + \sum_{j=1}^{8} \beta_j x_{i,j} \end{aligned} \tag{2.2}$$

where y_i=1 when a student earns grade C or above on their first attempt in Math 107. There is a β_j for each of the scores, grade and age of the SAT, ACT, AccuPlacer, and Devm/Math 107. Interactions were not included in the imputation or analysis model due to the inability of MICE package to easily handle the interactions. The age of tests and coursework were categorical with levels recent (1) and not recent (0) with the same definitions given in the complete case deletion model. Also grades for Math/Devm 105 and tests scores were coded the same as in the complete case deletion model.

For the last step, the calculations for pooling imputed datasets were straightforward and developed by Rubin (1987). These formulas may be found in Rubin (1987) or van Buuren (2012).

3. Simulation Study Set up

The purpose of the simulation study was to investigate and choose the best method between complete case deletion and multiple imputation. The simulation study had two degrees of missingness (high and low) and two missing data patterns (MAR and MNAR).

We simulated n=1,000 samples of predictors $\boldsymbol{X} \sim MVN_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\mu = (500, 20, 100)$ and Σ is the covariance matrix, corresponding to variances $\sigma^2 = (5500, 100, 50)$ and correlations as follows

$$\begin{pmatrix} 1 & 0.9 & 0.9 \\ 0.9 & 1 & 0.9 \\ 0.9 & 0.9 & 1 \end{pmatrix}.$$

The response variable Y_i was a binary response from Bernoulli(π_i) where

$$\pi_i = \frac{e^{x_i^T \boldsymbol{\beta}}}{1 + e^{x_i^T \boldsymbol{\beta}}}$$

$x_i = (1, x_{i1}, x_{i2}, x_{i3})$, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^T = (-25, 0.03, 0.06, 0.09)^T$.

The simulation had missing data patterns MAR between X_1 and X_2 and MNAR on X_3. For MNAR, when $X_3 < 200$, there was a 0.5 chance of missingness on X_3 for the low degree grouping and 0.7 chance on the high degree group. On average the percent missing on X_3 is 50%, and 70% for the low and high degree percent missing, respectively.

For MAR condition between X_1 and X_2,

when $500 < X_1 < 580$ and $22 < X_2 < 28$ then

x_1 is missing with probability $\omega_{middling}$

x_2 is missing with probability $\alpha_{middling}$.

otherwise

x_1 is missing with probability ω

x_2 is missing with probability α

In order to achieve a low and big degree of missing values, the following settings were used.

	Low Degree	High Degree
$\omega_{middling}$	0.2	0.4
$\alpha_{middling}$	0.4	0.8
ω	0.3	0.4875
α	0.6	0.975

.

The above setting created on average about 28% missing for both X_1 and X_2 on the low degree setting and about 47% missing for both for the high degree.

The missing data methods used in the simulation were complete case deletion, and multiple imputation. For complete case deletion, the logistic regression model was similar to the student data method for complete case deletion (equation 2.1), so there is a separate logistic regression model for each of the predictors X_1, X_2, and X_3. For multiple imputation, the R package MICE was used, and a Bayesian normal linear regression was selected for imputation of all variables. For the high degree of missingness there was 20 iterations of the MICE algorithm, and 10 iterations for low degree of missingness. The number of imputed datasets was five for both situations. The logistic regression model for the multiple imputation method was similar to the student data method for multiple imputation (equation 2.2) by having each predictor in the same logistic regression.

The simulation study consisted of 1,000 datasets with sample size 1,000 and the methods described above. To compare the methods, the confidence interval coverage was examined. Confidence interval coverage (CIC) was the proportion of times the true value of a coefficient was contained in its confidence interval. We attempted to achieve a significance level of 0.05, so theoretically, 95% of the generated confidence intervals should contain the true values of the β's. Also bias and

mean square error (MSE) were used for comparison; definitions are given below,

$$\begin{aligned}\bar{\beta} =& \frac{\sum_{i=1}^{k} \hat{\beta}_i}{k} \\ \text{Bias} =& \bar{\beta} - \beta \\ \text{MSE} =& \text{Var}(\hat{\beta}_i) + \text{Bias}^2\end{aligned}$$

where k=1,000 is the number of simulated datasets.

4. Results

4.1 Description of Missing values

Among the 4-year degree seeking students who first attempted Math 107, approximately 63.1% of the 4,793 students had received a final grade in the developmental course Math/Devm 105. The test scores generally had a high degree of missing data; 48.7% of students had either not taken or didn't submit SAT Math scores. Also, the ACT and AccuPlacer tests had high percentages of missing with 73.6%, and 61.7%, respectively.

There were about 15 combinations of missing and observed values; these are depicted in Figure B.2 in appendix B. About 56% of students had one of four combinations of observed and missing values: either (1) developmental math alone, (2) SAT score alone, (3) developmental math with SAT score or (4) developmental math with AccuPlacer score. The most abundant (16.4%) among the 4,793 students was "developmental math alone", and the second most abundant was SAT score alone (14%).

4.2 Complete Case Deletion

The complete case deletion model (equation 2.1) for Devm/Math 105 had statistically significant interaction between grade and age (P-value < 0.05), and it was reasonable to keep both main effect terms for grade and age (table 4.1). The full model had a AIC value of 3741.28, which was lower then that of the model with just an intercept, 3969.2. When Math/Devm 105 was taken recently (<2 years),

the estimated odds of passing Math 107 on the first attempt are multiplied between 1.808 and 2.179 for every one-unit increase in Math/Devm 105 grade, with significance level 0.05.

TABLE 4.1: Final Complete Case Deletion Models (equation 2.1). Only significant coefficients are shown here. For full model coefficients that show insignificant terms, see appendix C.

Model Term	Value	Std. Error	Wald Chi-Square	P-values*
Math/Devm 105 (n=3,023)				
Intercept	-0.1186	0.4808	0.0608	0.8052
Grade	0.1735	0.1246	1.9368	0.1640
Age (baseline: not recent (0))	-2.0037	0.5163	15.0624	0.0001
Grade*Age	0.5121	0.1334	14.7410	0.0001
SAT (n=2,460)				
Intercept	-1.1956	0.2879	17.2440	<.0001
Test Score	0.00339	0.000566	35.9095	<.0001
ACT (n=1,264)				
Intercept	-1.5469	0.3404	20.6472	<.0001
Test Score	0.0929	0.0159	34.2083	<.0001
AccuPlacer (n=1,834)				
Intercept	0.2759	0.1494	3.4071	0.0649
Test Score	0.0141	0.00369	14.6482	0.0001

* P-values are used to determine whether each term is significantly different than 0.

For the complete case deletion model (equation 2.1) that included students who had a SAT score, only the coefficient for the test score was statistically significant (P-value < 0.05). The coefficients in table 4.1 result after dropping the interaction and test age. The AIC for the model with only SAT Score and intercept was 3217.63, which, was lower than the AIC for the intercept only model with 3252.3. The estimated odds of first attempt pass in Math 107 for students who had a SAT score multiply between 1.257 and 1.569 for each 100-unit increase in SAT score. Or put another way, a 100-unit increase in SAT score has at least a 25.7% and at

most 56.9% increase in the odds of passing Math 107. Students who took the SAT test and scored 500 had between 60.3% and 64.2% probability of passing Math 107 on the first attempt.

Results for the complete case deletion model (equation 2.1) for ACT showed the interaction between test score and age was not statistically significant (P-value $>$ 0.05) and the same was true for the main term of test age. The model with the lowest AIC was one that included ACT Score and the intercept (AIC=1664.45), while the intercept only model had AIC of 1697.9. The ACT math score was significant (P-value $<$ 0.05) and the estimated odds of passing Math 107 on the first attempt lie between 1.362 and 1.895 for each 5-unit increase in ACT score (table 4.1). The probability of passing Math 107 on the first attempt for a student with a ACT math score of 20 was between 54.8% and 60.5%. The significance level was 0.05 for both confidence intervals.

Finally the complete case deletion model (equation 2.1) that included students with AccuPlacer scores showed a similar pattern of significance as the SAT and ACT models. AccuPlacer college level score and age interaction were not significant and neither was the main effect term for age of test (P-value $<$ 0.05). The test score was significant (table 4.1) and the model including the intercept and test score coefficient had a better AIC versus the intercept only model (2245.86 vs 2259.09). For each 20-unit increase in AccuPlacer score, the estimated odds of first-attempt pass Math 107 lie between 1.148 and 1.534 (table 4.1). Those students who scored 46 on the AccuPlacer test had between 69.9% and 73.% chance of passing Math 107 on the first attempt.

4.3 Multiple Imputation

The results from multiple imputation (equation 2.2) showed that Devm/Math 105 grade was significant (P-value $<$ 0.05) and the test scores were statistically insignificant except for ACT scores (table 4.2). We may infer that a one letter grade increase in Devm/Math 105 grade has between a 60% and 92% increase in the estimated odds of a student passing Math 107 on the first attempt, with significance level 0.05.

TABLE 4.2: Final Multiple Imputation Model. The multiple imputation had 30 iterations, and 10 imputed data sets. (n=4,793)

Model Term	Estimate	Std. Error	t-stat	df	P-values
Intercept	-2.932	0.334	-8.771	25.027	<0.0001
Math 105 grade	0.563	0.045	12.503	57.890	<0.0001
ACT Score	0.059	0.015	3.822	18.411	0.001

4.4 Simulation Study

Only results for $\hat{\beta}_1$ are shown; the β_2 and β_3 estimators showed similar results. When the degree of missingness is low, multiple imputation had the least bias (-0.0007) compared to complete case (0.0139) and the MSE was smaller (0.00003) with the multiple imputation method, see table 4.3. The same pattern exists between the methods under a high degree of missingness.

When there is a low degree of missingness, about 95.6% of the 1,000 confidence intervals from the multiple imputation method contained the true value of $\beta_1 = 0.03$ (table 4.3). The multiple imputation method met the theoretical coverage (95%) under a low and high degree of missingness. However, the complete case deletion method had poor coverage under both degrees of missingness.

TABLE 4.3: Comparison of β_1 estimate with high and low missingness along with two missing data methods. Complete case (CC) for just X_1, and multiple imputation (MI). CIC stands for confidence interval coverage. The true value for β_1 was 0.03. (simulation iterations: 1,000)

	Degree of Missing							
	Low				High			
Method	Bias	MSE	SE($\hat{\beta}$)	CIC	Bias	MSE	SE($\hat{\beta}$)	CIC
MI	-0.0007	0.00003	0.0051	0.956	-0.00078	0.00007	0.009	0.967
CC	0.0139	0.0002	0.0033	0.005	0.0138	0.0002	0.0039	0.014

5. Conclusions

Performance of missing data methods is dependent on the three types of missingness: MCAR, MAR and MNAR. In the simulation study with MAR, MNAR, high correlation and high missingness, the method of multiple imputation produced better confidence interval coverage (table 4.3) compared to complete case deletion. The simulation study does present some evidence that the complete case method produces terrible confidence interval coverage (table 4.3) under a missing data pattern involving MNAR and MAR. Future studies should consider even higher degrees of missingness and low correlation.

The missing data mechanism behind the Math 107 dataset is likely MAR or MNAR. The best method for the student data would then be the multiple imputation method since the evidence from the simulation suggests that confidence interval coverage is dramatically better for multiple imputation, and multiple imputation had the least biased estimators (table 4.3) under MAR and MNAR situations. We should be cautious and note that the simulation was for a specific set up and perhaps should not be overgeneralized.

For the student data, the age of tests were not significant under both missing data methods. The test scores were significant predictors of student success when using the complete case method (table 4.1), yet with the multiple imputation method the test scores were insignificant (table 4.2). One disadvantage of the multiple imputation method is the differences in parameter estimators after re-running imputation. This reflects the uncertainly of what value to impute (van Buuren, 2012). The ACT score was sometimes significant (P-value < 0.05) and others times not significant. The high percent of missing information among students (73%) for a ACT score is likely the reason for this issue.

It is obvious from the two missing data methods that the developmental course Devm/Math 105 grade was an important explanatory variable for predicting the

success of first attempts in Math 107 (table 4.1, table 4.2). A college course might offer more information about a student's ability to pass another college course. An interaction between the Devm/Math 105 grade and age would have been interesting to include in the analysis using the multiple imputation method, and future studies may want to investigate this.

Other explanatory variables for success in Math 107 exist such as high school math coursework that were not available for effective analysis. Future studies should consider the use of ALEKS, and high school math courses as predictors of student success. It is important to note that models are a simplification and there is likely not a correct model. The model presented here offers insight on whether placement tests and a specific developmental math course are predictors of success in Math 107; it illustrates the use of missing data techniques which is common to many data sets.

Appendix A

Math 107 Prerequisites

Table A.1: Cut off Scores For Math 107

	ACT Math	SAT Math	AccuPlacer College Level Math	COMPASS College Algebra	ASSET College Algebra
UAA	22-25	520-589	50-59	NA	NA
UAF	23-27	530-600	50-89	50-55	41-55
UAS	NA	NA	63-84	NA	NA

Table A.2: Recommended Age of Prerequisites

	Placement Test	Math/Devm 105
UAA	Within last two years	
UAF	One calendar year	Two calendar years
UAS		

TABLE A.3: Math 107 Course Description and Prerequisites by University

University	Description and Prerequisites
UA Anchorage	MATH A107 College Algebra 4 Cr Contact Hours: 4 + 0 Prerequisite: Math A105 with minimum grade of C. Registration Restrictions: If prerequisite is not satisfied, appropriate SAT or ACT scores or approved UAA Placement Test required. Courses Attributes: UAA GER Quantitative Skill Requirement. Special Note: A student may apply no more than 7 credits from any combination of Math A107, Math A108, and Math A109 toward the graduation requirements for any baccalaureate degree. Covers equation and inequalities, function theory, solution of equations, greater than first degree, matrices, determinants, systems of equations, and inequalities, exponential, and logarithmic functions, graphs and equations of conic sections, binomial theorem and sequences and series includes applications of all theses topics.
UA Fairbanks	Math F107X Functions for calculus (m) Contact Hours: 4 + 0 A study of algebraic, logarithmic and exponential functions; sequences and series; conic sections; and as time allows system of equation, matrices and counting methods. A brief review of basic algebra in the first week prepares students for rigor expected. The primary purpose of this course, in conjunction with Math F108, is to prepare students for calculus. Note: Credit may be earned for taking Math F107 or Math F161X, but not for both. Also available via eLearning and Distance Education. Prerequisites: DEVM F105 or DEVM F106 with a grade of B (3.0) or higher; or two years of high school algebra and Math F107X placement or higher. (4.5+0)
UA Southeast	Math S107 College Algebra 4 Credits (3+0) GER A detailed study of linear quadratic, rational, radical, exponential, logarithmic functions; operations on and applications of these functions, and select topics from algebra. Prerequisite: Math 105 with a C (2.00) or higher or placement test.

Appendix B

Figures and Graphs

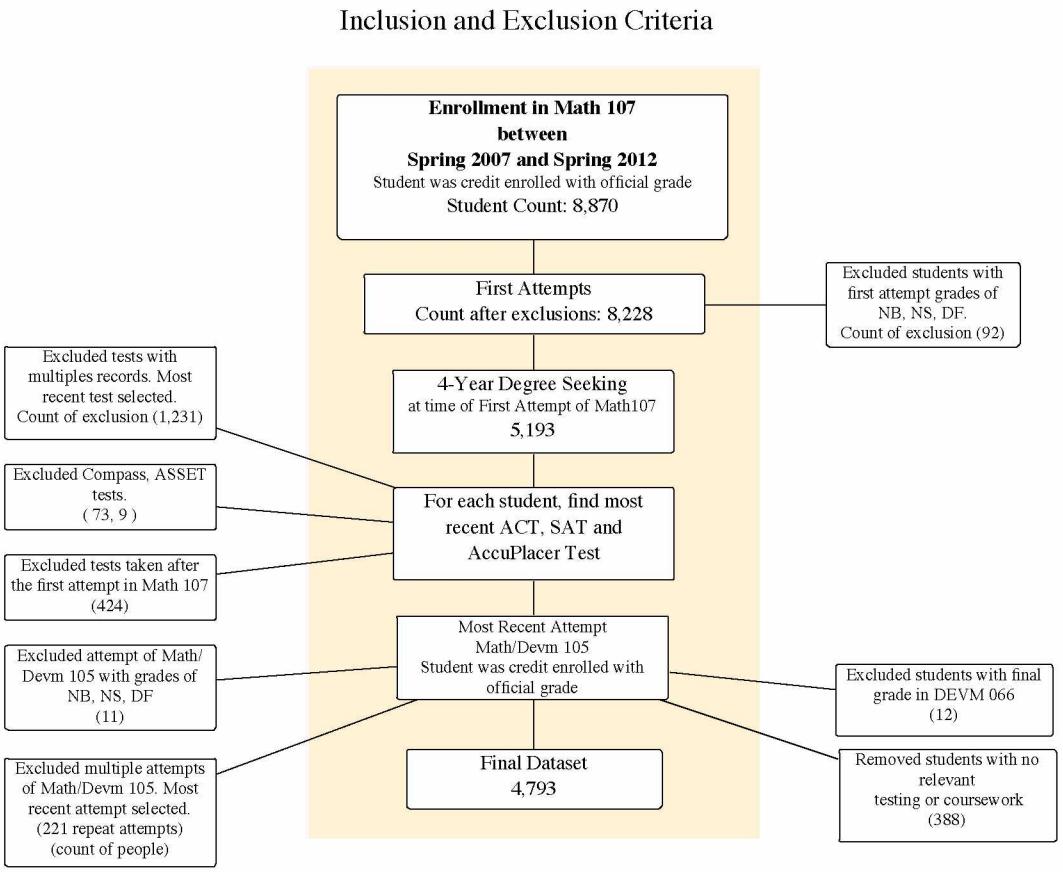

FIGURE B.1: Flow Chart of Inclusion and Exclusion Criteria for the study.

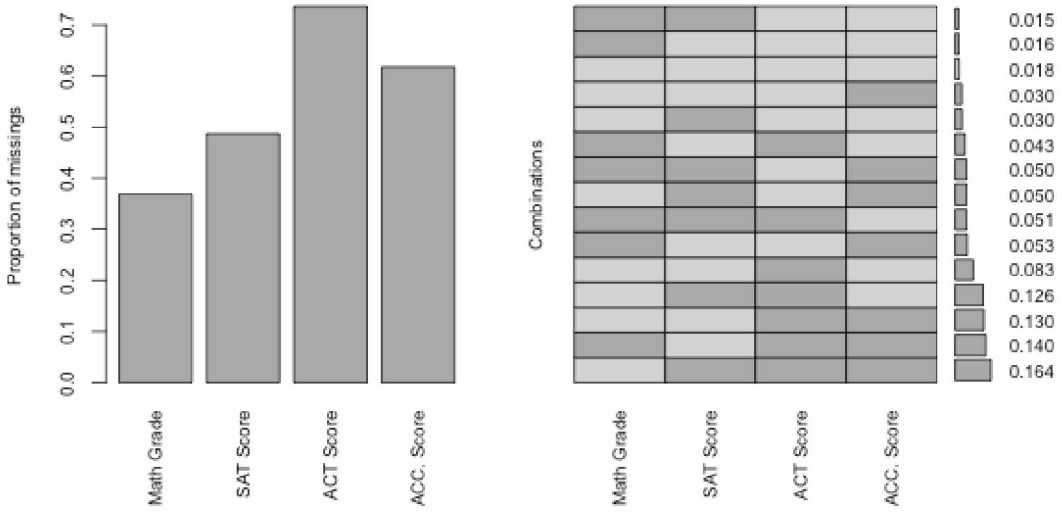

FIGURE B.2: Missing data combinations for students who first attempted Math 107 between fiscal years 2007 and 2012 (n=4,793). Light grey squares represent observed data and dark grey squares represent missing values. The numbers along the right side of the matrix graph represents the proportions of students with the combination over the total sample size. The math grade column refers to Devm/Math 105.

Appendix C

More Results

TABLE C.1: Full Complete Case Deletion Model.

Model Term	Value	Std. Error	Wald Chi-Square	p-values*
Math/Devm 105 (n=3,023)				
Intercept	-0.1186	0.4808	0.0608	0.8052
Grade	0.1735	0.1246	1.9368	0.1640
Age	-2.0037	0.5163	15.0624	0.0001
Grade*Age	0.5121	0.1334	14.7410	0.0001
SAT (n=2,460)				
Intercept	-1.0727	0.3804	7.9522	0.0048
Test Score	0.00302	0.000773	15.2068	¡.0001
Test Age	-0.1304	0.5931	0.0484	0.8259
Age*Score	0.000482	0.00117	0.1704	0.6798
ACT (n=1,264)				
Intercept	-1.4432	0.4973	8.4208	0.0037
Test Score	0.0900	0.0247	13.3098	0.0003
Test Age	-0.3171	0.7052	0.2022	0.6530
Age*Score	0.0112	0.0334	0.1131	0.7367
AccuPlacer (n=1,834)				
Intercept	-0.1087	0.2767	0.1543	0.6944
Test Score	0.0230	0.00782	8.6341	0.0033
Test Age	0.5802	0.3320	3.0543	0.0805
Age*Score	-0.0128	0.00894	2.0627	0.1509

* p-values are used to determine whether each term is significantly different than 0.

TABLE C.2: Full Multiple Imputation Model. ACC stands for AccuPlacer and m105 stands for Devm/Math 105.

Model Term	Estimate	Std. Error	t-stat	df	p-values*
Intercept	-2.912	0.379	-7.678	41.070	0.000
m105 grade	0.566	0.046	12.294	54.053	0.000
m105 age	-0.118	0.222	-0.531	24.654	0.600
SAT score	0.00015	0.001	0.112	14.581	0.913
SAT age	0.259	0.207	1.254	23.774	0.222
ACT score	0.053	0.031	1.734	13.190	0.106
ACT age	-0.292	0.235	-1.242	19.340	0.229
ACC score	0.005	0.008	0.617	11.450	0.549
ACC age	-0.037	0.111	-0.334	30.656	0.741

* p-values are used to determine whether each term is significantly different than 0.

Appendix D

R Code for Simulation

```
B <- 1000 # Number of Simulation Iterations
m <- 5 # Number of IMPUTED datasets
tmax <- 10 #20 of high degree of missing and 10 of low degree of missing
# For Creating Missingness MAR#
# Adjust porportions for more or less missingness #
# situation 1: low degree: 0.2, 0.4 AND 0.3, 0.6 # #
# situation 2: high degree: 0.4, 0.8 AND 0.4875, 0.975 #
x1middling_prob_cut_off  <- 0.2
x2middling_prob_cut_off <- 0.4
x1nonmid_prob_cut_off <- 0.3
x2nonmid_prob_cut_off <- 0.6
# For Creating Missingness MNAR#
x3_prob_cut_off <- 0.5 #low 0.5 and high 0.7

n  <-  1000   # sample size
rr  <-  matrix( c(1.0, 0.9, 0.9, 0.9, 1, 0.9, 0.9, 0.9, 1.0 ), nrow=3,ncol=3)
pp <- nrow(rr)
 x1 <- rep(NA,n)
 x2 <- rep(NA,n)
 x3 <- rep(NA,n)
 true.pi <- rep(NA,n)
 lin <- rep(NA,n)
 yy <- rep(NA,n)
 for( i in 1:n ) {
```

```
    dd <- mvrnorm( 1, rep(0,pp), rr )
    # mvrnorm(n = 1, mu, Sigma, tol = 1e-6, empirical = FALSE)
    x1[i]  <-  sqrt(5500)*dd[1]+500;  # similar to sat score#
    x2[i]  <-  sqrt(100)*dd[2]+20; # similar to act score#
    x3[i]  <-  sqrt(50)*dd[3]+100;
  }
x1 <- round(x1)
x2 <- round(x2)
x3 <- round(x3)
# linear regression part#
lin     <- -25 + 0.03*x1 + 0.06*x2 + 0.09*x3 ;
true.pi <-  exp(lin)/ (1+exp(lin));
yy      <-  rbinom( n, 1, true.pi )

############### -- create missingness -- #################
middling.x1 <- function(x1) # SAT-like
{
return ( (x1>=500) & (x1<=580) )
}
middling.x2 <- function(x2) # ACT-like
{
  return ( (x2>=22) & (x2<=28) )
}
mx1 <- x1
mx2 <- x2

which.middling <- (middling.x2(x2) & middling.x1(x1))
which.not.middling <- !which.middling
which.middling <- which.middling * (1:n)
which.not.middling <- which.not.middling * (1:n)
which.middling <- which.middling[ which.middling != 0 ]
which.not.middling <- which.not.middling[ which.not.middling != 0 ]

for( j in which.middling ) {
###          w.p. 0.2, toss x1, retain x2
###          w.p. 0.2, toss x2, retain x1
```

```
###          w.p. 0.6, keep both

  cointoss <- runif(1)
  if( cointoss < x1middling_prob_cut_off ) {
    mx1[ j ] <- NA
  } else if (cointoss < x2middling_prob_cut_off ) {
    mx2[ j ] <- NA
  }
}

for( j in which.not.middling ) {
###          w.p. 0.4, toss x1, retain x2
###          w.p. 0.4, toss x2, retain x1
###          w.p. 0.2, keep both

  cointoss <- runif(1)
  if( cointoss < x1nonmid_prob_cut_off ) {
    mx1[ j ] <- NA
  } else if (cointoss < x2nonmid_prob_cut_off ) {
    mx2[ j ] <- NA
  }
}

mx3 <- x3
for( j in 1:length(x3) ) {
cointoss <- runif(1)
  if( x3[j] < 200 & cointoss < x3_prob_cut_off ) {
    mx3[ j ] <- NA
  }
}
############### -- end of create missingness -- #################

mi <- data.frame(yy, mx1, mx2, mx3)

####### Multiple Imputation ######
imp <- mice(mi, method=c("","norm","norm","norm"), m=m, maxit=tmax, print=F)
```

```
fit <- with(imp, glm( yy ~ mx1 + mx2 + mx3, family=binomial(logit)))
pfit <- pool(fit)

####### Complete Case ######
separate_x1 <- glm( yy ~ mx1, family=binomial(logit), data=mi)
sx1.summary <- summary(separate_x1)

separate_x2 <- glm( yy ~ mx2, family=binomial(logit), data=mi)
sx2.summary <- summary(separate_x2)

separate_x3 <- glm( yy ~ mx3, family=binomial(logit), data=mi)
sx3.summary <- summary(separate_x3)
```

References

Belfield, C. and Crosta, P. (2012). *Predicting Success in College: The Importance Of Placement Tests and High School Transcripts. Working Paper No. 42.* New York, NY: Community College Research Center, Teachers College, Columbia University.

Hiss C. W., and Franks V. W. (2014). *Defining promise: optional standardized testing policies in American college and university admissions.* Arlington, VA: National Association for College Admission Counseling.

Baraldi, A. N. and Enders C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology*, 48(1):5-37.

Collins, L. M., Schafer, J. L., and Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 6(3):330-351.

Rubin, D. B. (1976). Inference and missing data. *Biometrika.* 63(3):581-590.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley & Sons.

Schafer, J. L. and Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7(2):147-177.

van Buuren, S. (2012). *Flexible Imputation of Missing Data.* Florida: Taylor & Francis Group.